

7

Designing Incentive Systems for Schools

Derek Neal

Much debate concerning the design of performance incentives in education centers on specific psychometric challenges. Advocates of the use of performance incentives in education often argue that student test scores provide objective measures of school output, but their opponents raise concerns about the breadth and reliability of assessments, the alignment of assessments with curriculum, and the potential for schools to manipulate assessment results through various forms of coaching or even outright cheating. In sum, many doubt that school systems can or will construct student assessments that truly form a basis for measuring and rewarding educational performance.¹ Although these psychometric concerns are first order, I argue that policymakers and researchers must pay more attention to challenges that would remain in a world with perfect assessments.

Assume for a moment that the only mission of schools is to foster the math skills associated with a particular curriculum. Furthermore, assume that policy-

I thank the participants at the National Center on Performance Incentives conference for useful comments. I thank Matthew Springer for detailed suggestions. I thank Margarita Klerkh and Richard Olson for research assistance. I thank Gadi Barlevy, Edward Haertel, and Canice Prendergast for useful discussions. I thank the Seale Freedom Trust for research support. I also thank Lindy and Michael Keiser for their support through a gift to the University of Chicago's Committee on Education.

makers in this setting possess an ideal instrument for assessing math skill and are able to make assessments of every student at the beginning and end of each school year. Even these ideal assessments do not provide the information policymakers need to rank schools according to their performance.

If a factory produces five hundred widgets today, we know that the value of this output is five hundred times the price of a widget. If Johnny's math scale score rises from 140 to 145, we may conclude that Johnny's expected number of correct answers, in a setting that requires him to try all the items in a specific domain, has increased by 5. However, we do not know what this increase in expected correct answers is worth to Johnny or to society. In addition, we do not know whether a 5-point increase would have been worth more to society if Johnny had begun the school year with a baseline score of 130 or 150 instead of 140. Finally, because Johnny may receive tutoring and support from his parents as well as his teachers, we cannot straightforwardly determine what portion of Johnny's score increase should be credited to his school rather than his family.

Education is not the only field in which it is difficult to attach dollar values to the marginal contribution of a given worker or a group of workers who function as a production unit. Private firms that face these measurement issues often abandon the task of trying to produce cardinal measures of output for individual workers or teams. Instead, firms take on the more manageable task of forming performance rankings among workers or groups of workers and then deliver bonuses, raises, and promotions as a function of these rankings.²

However, the task of constructing performance rankings in public education differs from the task of constructing performance rankings in most private firms because there is no clear way, *a priori*, to collapse the multiple dimensions of school performance into a single index. Private sector firms may not be able to precisely measure the contribution of a worker or group of workers to overall profits, but firms know that this is the criterion by which they seek to rank performance. In public education, policymakers must begin the process of designing incentive systems by developing a clear definition of performance and then pay close attention to the mapping between this definition and the performance ranking procedures they adopt.

Some scholars suggest that because the potential benefits of cooperation among teachers are large relative to the costs of cooperation, incentive systems in education should provide rewards and punishments at the school level. Others have touted the benefits of allowing individual schools or organizations that manage groups of schools to compete not only in academic performance contests but also in the labor market for teachers. Systems that assign reward pay at

the school level but allow each school to allocate resources among teachers according to its own personnel policies foster competition in the market for teachers that may speed the rate of social learning about the best ways to hire, mentor, and motivate teachers.

Most of the analyses offered here rest on the implicit assumption that there exists a benevolent education authority that faithfully represents the interests of taxpayers, but it may be the case that the public provision of education invites political corruption that contaminates the design of incentive systems. This observation raises the possibility that voucher systems serve as complements to rather than substitutes for incentive pay and accountability systems.

The Limits of Performance Statistics

Private firms have the ability to hand out bonuses, promotions, and other forms of reward pay based not only on objective information and procedures but also on the subjective evaluations of owners or the managers who work for them. This arrangement is possible because workers know that owners are losing their own money when firms fail to retain, motivate, and promote their best employees. However, there are no residual claimants in government agencies, and officials who run public organizations may suffer no harm if they hand out bonuses and reward pay to their friends and family instead of to those who are most deserving. This feature of public agencies generates demands by public employees that performance incentive systems in government tie rewards and punishments to objective performance measures, and these performance statistics are often reported to the public.

In 1976 Donald Campbell made the following observation concerning government statistics that is often referred to as Campbell's law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."³ Campbell's law in fact makes two assertions: First, when governments attach important stakes to specific government statistics, actors within governments often face incentives to engage in activities that improve these statistics without actually improving the conditions that the statistics are meant to monitor. Such activities corrupt the statistics in question because the statistical improvements induced by the activities do not coincide with real improvements in welfare. Second, the same activities that corrupt performance statistics may actually cause direct harm.

Campbell provides numerous examples of this phenomenon, and in chapter 5 in this volume, Richard Rothstein provides more detail concerning Campbell's

observations and related observations from several different fields. Because Rothstein's summary of existing evidence suggests that Campbell's law may be an appropriate label, education policymakers should be wary of performance incentive or accountability systems that rely heavily on performance statistics. Workers change their behavior in response to the adoption of any particular performance measurement system, and these responses often compromise the value of the performance measures in question.

Modern economists typically use Bengt Holmstrom and Paul Milgrom's multitasking model to organize their analyses of the phenomena that Campbell describes. Holmstrom and Milgrom built their model to explain why private firms often choose not to attach incentives to performance statistics even when they have access to statistics that are highly correlated with actual performance. Their insights concerning the settings in which private firms are reluctant to attach high stakes to performance measures help us understand why Campbell drew such pessimistic conclusions about the use of performance statistics in government.⁴

In the multitasking model, agents may take many different actions at work, and by assumption, employers have two tools for directing the efforts of their workers. First, firms may pay the costs required to monitor their workers' actions directly. Second, firms may link worker pay to performance statistics. These statistics are noisy signals of worker outputs, and the key assumption is that the relationships between worker actions and measured output are not the same as the relationships between worker actions and actual output. Some actions have a greater impact on measured output than actual output, while the reverse is true for other actions.

Advocates of recent trends in education reform hope that high-stakes assessments will prompt teachers to allocate more effort toward activities like teaching math and less toward activities that amount to leisure for teachers and extra recess for students, and any fair assessment of test-based accountability programs would likely conclude that accountability systems do create these types of changes in effort allocation.⁵ However, the logic of the multitasking model suggests that other reallocations should also be expected.

Everyone knows the saying, "You get what you pay for," but the multitasking model takes this line of reasoning a step further. Because effort is costly, if firms pay for performance as measured by a statistic, workers will not only allocate more effort to the actions that have the greatest impact on the statistic in question but also allocate less effort to actions that do not have large direct impacts on this statistic. Furthermore, this reallocation will occur even if it means that less effort will be devoted to actions that have large positive impacts on actual output. In educa-

tion, these reallocations may involve teachers' spending less time on activities that foster creativity, problem-solving skills, the ability to work well in teams, or other important skills that are not directly assessed. Thus even if teachers put forth more total effort following the introduction of assessment-based incentive pay, these types of reallocations may leave students worse off.⁶

Campbell's empirical observations combined with the insights of the multitasking model are warning signs for those who wish to design incentive pay systems for public schools. However, considering the design challenges inherent in constructing incentive pay systems for educators, the corruption of assessments is only the tip of a large iceberg.

Necessary Ingredients

Here, I discuss the task of ranking educators in an environment with ideal assessments. Incentive pay systems for educators require two components. First, these systems require a method for ranking schools or teachers according to performance. Second, they require the assignment of specific rewards and penalties to the various performance ranks that schools or teachers may receive.

Assume the following are true:

- There are exactly K skills that schools are supposed to foster.
- Each school has N students.
- There exists an assessment for each of the K skills.
- Each of the K assessments has perfect reliability.
- Neither students nor teachers can corrupt the assessment results.
- Variation in achievement growth among students is determined entirely by school performance.

The assumption of no corruption implies that the only way schools can enhance their students' test scores is to engage in activities that create more skill among their students. The final assumption implies that policymakers can isolate the contribution of schools to student skill development.

This ideal setting brings to the forefront a fundamental issue that must be settled before the process of designing incentive systems for schools can begin. To design a system that rewards performance, performance must first be defined. Note that, at a point in time, all the assessment information for the students in a given school can be placed in an $N \times K$ matrix. Each row contains all the assessment results for a particular student, and each column contains the results of a particular assessment for all the students in that school. If we index schools by

ners know the criterion they are supposed to use. They are supposed to rank associates based on their best guesses concerning how much each associate could add to the total value of the partnership. However, if the superintendent of a large school district or even a state decides to rank schools or teachers according to their performance, she or he must first construct a definition of performance.

Any sensible method of constructing performance rankings in education must be guided by three principles that are all variations on the same theme: spell out priorities; clearly map priorities and procedures; and define sensible comparison sets. A coherent definition of performance must serve as an anchor for the procedures used to construct performance rankings.

Spelling Out Priorities

First, the documents describing any accountability or incentive pay system should spell out the priorities of policymakers. These documents should clearly delineate the types of achievement that the system is intended to foster and, to the extent possible, should explore how policymakers view the relative importance of achievement in various subjects or by various types of students. Thus, policymakers should begin by formulating clear answers to questions like the following:

—Is progress in reading more valuable than progress in math or civics, and if so, how much?

—Is it more socially valuable to bring a disadvantaged student closer to grade level than to bring a gifted student closer to her or his full potential, and if so, how much?

—What are the relative values of noncognitive traits, like persistence, and cognitive skills?

Schools are supposed to simultaneously foster many skills in hundreds of students at the same time. Without clear answers to these questions and many others, the task of objectively ranking the overall performance of any two schools is a hopeless endeavor.⁷

Mapping Priorities and Procedures

Second, the mapping between the policy priorities that define an incentive system for educators and the procedures used to create performance rankings for schools and teachers should be clear and precise. This step is challenging, but those who design and implement incentive systems risk failure if they do not devote enough attention to this essential task.

$s = 1, 2, \dots, S$, we can define a set of S matrixes as $X = (X^1, X^2, \dots, X^S)$. Each matrix is $N \times K$, and together these matrixes contain all skill assessments for all students in all schools at a point in time. For simplicity, I assume that these measurements are taken on the first day of school. Next, I define X' as the collection of measurements $X' = (X^1, X^2, \dots, X^S)$ taken among the same students on the last day of the same school year. Given that society began the school year at X , how does society evaluate the relative values of ending the year at any one of the billions of possible X' outcomes? Furthermore, if I take the matrixes of test scores from the beginning and end of the school year for any two schools, how do I use these four matrixes to decide which school performed better?

In a truly perfect world, an incredibly skilled team of econometricians possessing the largest research grant in the history of social science research would have already devised a method for estimating the social value (in dollars) of moving the students in school s from any X^s to any X'^s , and given this method, it would be easy to design incentives for educators. Education authorities could simply allow competing school districts or school management companies to bid for the opportunity to operate schools in given locations and then pay each of these entities a lump sum at the end of the year equal to the social value of the change in human capital among all of its students minus the total amount bid for the right to operate schools.

This simple approach is not possible in education, and this idealized setting shows that the central reason is orthogonal to common observations concerning the difficulty of accurately assessing all the skills produced in schools. Even if policymakers possessed measures of all skills produced in schools, and these measures were reliable and expressed on interval scales, policymakers would still have no idea how to value various improvements on these scales in monetary terms.

Even psychometrically perfect assessments provide no rational basis for constructing pay-for-performance systems that look like piece rate or commission systems; furthermore, they do not provide the information required to simply rank schools or teachers according to performance. Because school output is multidimensional, that is, there are $N \times K$ outcomes at each point in time in each school, it is not clear a priori how one collapses this information into a one-dimensional performance ranking for schools or teachers.

Many owners and managers in the private sector also operate in environments that do not permit them to assign a dollar value to the marginal contributions of each of their employees, but the task of constructing performance rankings is probably more complicated in education than in these private firms. If the partners in an accounting firm sit down to create a ranking of their associates, all part-

Consider the No Child Left Behind Act of 2001 (NCLB) as an example. The language of the act, beginning with its title, signals that addressing the educational needs of the most academically disadvantaged is a high priority. However, Derek Neal and Diane Schanzenbach argue that, in states that measure school performance by counting the number of students who score above a statewide proficiency standard, the levels of the proficiency standards on various assessments determine which students are most pivotal for a school's performance rating. Students who test below the proficiency standard but are able to achieve the standard given modest interventions are the students whose achievement gains matter most in determining their school's status under No Child Left Behind. Thus even though the rhetoric surrounding NCLB highlights the need to improve outcomes among our most disadvantaged students, NCLB implicitly places greater social value on the learning of students in the center of the achievement distribution than on that of students who are currently far below their state's proficiency standard.⁸

In states that use value added systems to measure school or teacher performance, choices of scales for the exams combined with choices concerning how to weight improvements that occur in different ranges of the test-score distribution determine the rewards that schools receive for directing attention to different students and different subjects, but policymakers often fail to offer a rigorous justification for these choices. A concrete example helps make this point clear. In the 2006–07 school year, Florida implemented the Special Teachers Are Rewarded (STAR) incentive system. An important component of the STAR program involved assigning performance points to teachers based on their students' gains on standardized tests using the value table method. Table 7-1 is an example of a value table. The Florida Department of Education offered this table as a model for how points should be assigned to teachers under STAR based on their students' reading outcomes.

Table 7-1. *Elementary Reading Value*

	Year-1 level, 2005	Year-2 level, 2006					
		1a	1b	2	3	4	5
1a		0	100	455	550	675	725
1b		-50	50	145	265	340	500
2		-100	-50	125	205	245	350
3		-175	-100	-90	170	210	250
4		-200	-150	-140	-75	195	215
5		-250	-200	-160	-125	25	210

There are six levels of reading achievement for students in Florida elementary schools, and the table specifies points associated with each of the thirty-six possible student transitions. As the table indicates, if a student moves from level 2 to level 3 in one year, her or his teacher receives 205 points. However, the teacher of another student who moves from level 1b to level 2 receives only 145 points. The department of education intentionally awarded more points for improvements that are less common, but it is hard to see why these particular gradients are the right ones.

The additional reward for bringing a student past level 3 and up to level 5 in year 2 varies greatly depending on the baseline achievement level. The marginal reward is much greater if the student began at level 1b than at either levels 1a or 2. Why would this be the case? Shouldn't one expect that the value to society of bringing a child from level 3 to level 5 is roughly the same regardless of the child's identity? If Johnny began the year behind Sue but both Johnny and Sue are at the same reading level by January, is there any reason that society should value Johnny's learning during the spring more or less than Sue's?

Because the STAR proposal did not contain a detailed discussion of the relative social importance of different types of progress among different types of students, it would be easy to generate an equally plausible set of point allocations for the entries in table 7-1 that would imply notably different results in terms of which teachers are ranked among the top performers in their district. The STAR system and other systems that do not create clear ties between how performance is defined and how performance is measured inevitably yield performance rankings that lack credibility.

Defining Comparison Sets

Third, incentive systems should group schools according to the types of students and families they serve and then rank schools either relative to other schools that serve similar students or to a performance standard designed for such schools. Any attempt to create a single performance ranking over all schools in an entire state or large district necessarily encounters serious conceptual problems. When school A is ranked above school B, the implication is that school A performs better than school B. However, if the two schools are educating students from extremely different backgrounds, one must ask, "Better at what?"

In 2006 Hillsborough County, Florida, decided to participate in the STAR merit pay system. Although STAR's value table approach sought to place all teachers on a level playing field, the 2006–07 results in Hillsborough suggest that the STAR procedures generated performance rankings that overstated the true per-

formance of teachers in affluent schools relative to the those in disadvantaged schools. County officials moved quickly to modify the plan, and the revised plan involves schools' being placed in leagues according to their Title I status.⁹

The Hillsborough experience is not surprising given that the original plan sought to make performance comparisons among teachers who, in important respects, were not performing the same job.¹⁰ The tasks of defining and measuring job performance in education are necessarily complicated because educators perform complex jobs, but these tasks become meaningless when policymakers insist on making performance comparisons among persons who are not working with comparable students.

The gains that students make, in a given year, on any particular assessment scale reflect the interaction of their initial skill level and the quality of the instruction they receive. Thus data on students from two classrooms who began the year at widely different levels of achievement do not provide any information that allows one to directly compare the quality of instruction in the two classrooms. One can never rule out the possibility that students in one classroom simply began in a region of the scale where it is easier to make significant gains.

Auxiliary Benefits of Competition within Leagues

Some will worry that a system requiring schools to compete only against other schools that draw from similar student populations may do little to improve performance in disadvantaged communities because it may be possible for some schools to outperform most schools in disadvantaged communities without actually performing at an exceptionally high level. However, this line of reasoning does not take into account that teachers and principals might change where they choose to teach in response to such a system.

Imagine that there are ten different leagues in a state and that these leagues are defined by the early preparation and family backgrounds of entering students. If an "easy" league exists in which it is less costly to win reward pay or avoid sanctions, talented principals and teachers will face a strong incentive to move to a school in this league. More important, teachers and principals who are best suited to teaching in the schools that belong to that particular league face the strongest incentive to move.

Furthermore, in a system with league-specific tournaments, differences in reward pay across leagues can be used as an effective means of attracting the right teachers and principals to serve in disadvantaged communities. Those who respond to the extra reward pay are not only those who are willing to teach in dis-

advantaged communities but also those who are willing to bet that they know to do it successfully.

Finally, by using schools with observationally similar students to define the performance standard for any given school, one minimizes an important performance measurement problem that has been assumed away in the analyses presented so far. If Johnny's math score rises by ten points this year, it is hard to know what part of this gain should be attributed to the efforts of Johnny's teacher and what part to inputs Johnny receives outside school from parents, grandparents, or other adults.

To the extent that teachers and principals have information about the backgrounds of their children that are not reflected in the measures of preschool preparation or family background available to policymakers, it will not be possible to form perfect comparison groups for any school. However, to the greatest extent possible, whenever school A receives a better ranking than school B, this outcome should imply that school A performed better than school B and not simply that school A worked with more-advantaged students.

The state of California actually produces a performance ranking of each school within a set of schools that are similar in resources and the background of their students. Nevertheless, policymakers in California treat similar schools' rankings as simply "additional contextual information."¹¹ Neither the state accountability system nor the state implementation of NCLB attaches important reward sanctions to the ranking system's outcomes.

The federal government, in its implementation of NCLB, and numerous states continue to make the mistake of asserting that rewards and punishments for educators must be determined by measures of schools' performance relative to either statewide standards or all other schools in the state. Defenders of this approach argue that it is the only way to implement high achievement standards for all children, but this argument confuses two distinct uses of statistics.

Statistics that accurately indicate whether an organization is reaching its goals are not necessarily the statistics that organizations should employ in their incentive pay systems. If the object is to determine whether the children in a given state are reaching a minimum level of achievement as set by the state, then the performance of each student will need to be measured against a common standard that reflects this target. However, if assessment results are to be used as part of a set of personnel practices that rewards and punishes teachers and principals for their job performance, then comparisons must be made among persons who are working in comparable environments and thus doing comparable jobs.

Neither value added models nor growth models offer a way around this concern. The original Hillsborough approach sought to rank teachers using measures of achievement growth, and it still produced results that were not credible. If the baseline achievement distributions for two classrooms have little overlap, the data permit few direct comparisons between students who began the year at similar achievement levels but attended different classrooms. Although researchers or policymakers can always create models that produce estimates of how the teachers in these two classrooms are performing relative to one another, it is the modeling choices of analysts, not data on actual relative performance, that drive these estimates. Some will argue that Hillsborough simply chose the wrong growth model, but the county's real mistake was trying to make performance comparisons among teachers who were not working in comparable classrooms.

School versus Teacher Performance

Thus far this chapter has not drawn distinctions between incentive systems that operate at the school level and those at the teacher level and has often discussed incentive pay and accountability systems as if they operate at the school level. Nonetheless, the process of designing incentive systems in education requires that choices be made concerning the extent to which policymakers attach incentives to measures of overall school performance rather than of individual teacher performance.

Three different scenarios form interesting baselines. First, a districtwide or statewide system might link measured performance for individual teachers to teachers' pay and job security. Second, district or state policies might tie incentive pay for teachers to measures of how their schools or departments perform. Finally, a system might link all government performance incentives to school-level outcomes but allow those who run schools to adopt their own policies concerning how incentive payments at the school level are allocated among different teachers within schools. For a variety of reasons, the latter two approaches are likely preferable to the first.

Cooperation and Information Sharing

It seems reasonable to assume that the teachers in a school possess a great deal of information concerning how the performance of their peers could be improved. However, incentive systems that rely solely on rewards and punishments for individual teachers do not provide any motivation for teachers to share this valuable information with their peers. Thus even if an assessment-based system can accu-

rately show that teacher A is not performing as well as her peers, the system will not foster efficient improvement in teacher A's performance if teacher A is the only person affected by her performance.

For at least two reasons, an efficient system will provide incentives for teacher A's principal and peers to help her improve. First, they are likely to have the best information concerning how she might improve. Second, the costs of sharing this information, relative to the benefits, are often low. When one teacher shares with another teacher lessons learned from experience and experimentation, the time costs required to convey information may often be quite low relative to the benefits, and it takes little imagination to come up with numerous examples. Information concerning pedagogy, organization, or even the personalities and needs of particular students in the school may often be shared at low cost but to great benefit.¹²

Incentive systems based on measures of individual teacher performance not only provide no incentive for teachers to engage in this type of information sharing but may also provide clear incentives to withhold such information. Any system that makes performance comparisons among teachers working in the same school actually creates incentives for teachers to sabotage the performance of their peers. Although some may view this conclusion as far-fetched, economists point to this possibility as one reason that incentive systems used in the private sector are often designed to avoid direct competition among workers who are supposed to cooperate with one another in joint production activities.¹³

Some may argue that these undesirable effects can be avoided by having individual teachers compete against a fixed performance standard rather than one another. However, competition against fixed performance standards creates other problems. To begin, competition against a performance standard is competition against some historical notion of what was possible in the past in a particular type of classroom. This form of competition cannot require educators to perform at efficient levels unless standards are constantly revised to reflect what is possible in different school environments given new methods of pedagogy, instructional technologies, and other resources.¹⁴ Furthermore, this need for revision and updating creates opportunities for political forces to build low performance expectations into the system. Competitions that allow the possibility that everyone can be a winner invite mischief that lowers standards.

In contrast, when incentive systems involve direct competition among schools for reward pay, individual teachers have clear incentives to help their peers improve because they receive no reward pay unless their school performs better than other schools. Furthermore, if they have the freedom to hand out different

shares of their school's total reward pay based on their own aggregation of test-score outcomes and their subjective evaluations of each teacher, principals can build reputations for rewarding not only individual performance but also cooperation among teachers. Principals have strong incentives to pursue this course of action if their pay and job security depend on their overall performance rankings in their schools, and principals who follow this course strengthen incentives for teachers to help one another improve.

None of the above arguments against attaching incentive pay to measures of individual teacher performance deny that variation in individual teacher performance is an important factor in determining variation in student outcomes. Everyone who has ever been a student knows that some teachers are much better than others, and recent work by Steve Rivkin, Eric Hanushek, and John Kain provides clear evidence that this is the case. Identifying, training, and retaining talented teachers is key to running an effective school, and these tasks are too difficult to accomplish within systems that do not encourage all agents in a given school to use their information in ways that improve not only their individual performance but also the performance of others.¹⁵

An Easier Measurement Problem

Incentive pay systems based on school performance are also easier to implement than systems built around measures of individual teacher performance because it is difficult to measure differences in performance among teachers. The existing empirical literature provides clear evidence that teachers differ in efficiency but less clear evidence that statisticians can build reliable measures of teacher performance that form a credible basis for incentive pay. Several issues complicate the task of creating performance measures for individual teachers, especially statistical noise and classroom teacher assignments.

NOISE

Estimates of individual teacher effects for a given year are quite noisy when one attempts to include reasonable controls for student and classroom characteristics. Although a number of researchers have argued that a particular type of value added model can produce more reliable estimates of individual teacher effects by using multiple years of data, I do not see how a method that delivers precise estimates of teacher performance over periods of three to five years is useful as a basis for making personnel decisions and handing out reward pay.¹⁶

Most professionals in the private sector work in environments that involve some form of reward pay on at least an annual basis that comes in the form of

bonuses, raises, or profit sharing. Although decisions about promotions are made at less frequent intervals, promotion systems not only provide incentives for current effort but also affect the efficiency of the entire organization by allocating the most talented and productive people to positions in which success depends most heavily on talent and productivity. Performance measures for individual teachers derived from many years of data may be useful inputs for a tenure evaluation process, but they are not useful as a means of providing incentives for existing teachers, especially tenured ones, to provide efficient effort levels on a continuous basis.

IGNORING CLASSROOM ASSIGNMENTS

Jesse Rothstein highlights a second challenge for those who wish to use statistical methods to rank teachers based on their contribution to student achievement. Using North Carolina data, Rothstein shows that the identity of a student's teacher in a future grade helps predict performance in the current school year. This pattern is consistent with the hypothesis that the allocation of students to teachers within a school is driven, at least in part, by how individual students are progressing through school. Rothstein presents evidence that this sorting of students to teachers is driven not solely by fixed student characteristics but also by how the student develops academically over time, and he argues that estimated teacher effects based on methods that seek to control for this type of student tracking over time are not highly correlated with estimates from more standard models.¹⁷

Standard methods that researchers use to measure the relative performance of individual teachers rely on the assumption that given standard student background variables, the assignment of children to teachers is a random process and can therefore be ignored. However, the assignment of teachers to students within schools reflects a set of choices made by principals based on information that researchers cannot see. Some teachers excel at working with children who are naturally hard workers, while other teachers have a comparative advantage in working with kids who are struggling in school or at home. Thus when researchers assume that the assignments of teachers to students can be ignored, they are, in effect, assuming that principals systematically fail to do their jobs.

Ignorable assignment is still a challenge at the school level. Parents' choice of schools for their children most likely reveal information about unmeasured family characteristics that influence academic outcomes for their children. However, there are scenarios that make ignorable assignment at the school level much more palatable.

California already has a set of procedures that are designed to identify a comparison set of similar schools for any given school in California. In large states, it may be possible to form comparison sets that are not only homogeneous with respect to student characteristics but also geographically separated. Imagine a set of fifty elementary schools that serve as the comparison set for elementary school A. Assume that all fifty schools are similar to A with respect to early preparation for school and the demographic characteristics of students, and also assume that no student in any of these fifty schools lives within a one-hour commute of school A. That students in school A did not attend one of the fifty schools in this comparison set provides no information about school A or the comparison schools. The comparison schools were not realistic options for the students in school A. Furthermore, that students in the comparison schools did not attend school A is not informative about either school A or the comparison schools because school A was not an option for these students.

If such a comparison set is used to create a performance measure for school A, unmeasured factors at the community level may still create problems. However, there is no set of decisions facing parents, teachers, or principals that can be expected to directly generate correlations between these unobserved factors and the assignment of students to either school A or the schools in its comparison set.

The Value of Labor Market Competition

Assume that a state or large district allows independent companies and nonprofit organizations to bid for opportunities to manage public schools. Furthermore, imagine an incentive system that provides reward funds at the school level based on an index of school performance and also provides for the termination of an organization's management contract if the same index falls below a specified level. This index might be based entirely on assessment results or a combination of assessment results and the results of school inspections, parent surveys, and other information. Regardless, the key assumption is that the index is a reliable indicator of how a particular school performs relative to similar schools that serve students from the same backgrounds.

In addition, assume that the organizations that manage schools are responsible for distributing reward money to teachers and for designing their own policies and procedures for evaluating teachers, screening new hires, terminating existing teachers, granting tenure, and determining career ladders for teachers within their schools. Thus school management organizations compete with one another not

only in determining the educational practices used within their schools but also in developing and implementing the personnel policies and procedures that identify and retain the best teachers. Because the resources of these organizations are tied to their performance, they face clear incentives to select personnel policies that retain and reward the teachers who make the greatest contributions to overall school quality. Furthermore, as different organizations experiment with different management models, successful innovations will spread to other organizations and other schools.

This type of labor market competition among schools is almost never seen in the developed world. Although many European countries have education systems with voucher components that foster competition among schools for students, collective bargaining on a national or regional level sets most personnel policies for both private and public schools in these systems.¹⁸

The personnel economics literature describes many ways that private firms implement desirable performance incentive systems even in environments like education, in which it is impossible to precisely measure the marginal contributions of individual workers to the profits of firms. However, these papers usually describe incentive schemes that are possible only when firms know a great deal about both the preferences of their workers and the details of their production technologies.¹⁹ Economists justify this approach to characterizing what firms actually do by noting that competition among firms for talented workers moves the actual personnel policies of firms toward the efficient ones.²⁰ Inefficient policies waste resources either by paying too much for the effort that workers provide or by encouraging workers to provide effort that does not generate returns in excess of the incentive payments made to workers. Because firms that do not discover efficient ways to provide incentives for their workers waste resources and cannot compete over the long term with firms that do, competition in the product market enhances efficiency in the labor market.

For this reason, systems that promote competition among schools while allowing schools to compete for teachers by experimenting with different personnel policies offer greater promise than systems that impose a single set of incentive pay policies on all schools. Imagine that a state or district superintendent must design a single incentive pay system for an entire state or district. Even if possessing an ideal system for creating teacher performance rankings based on peer evaluations, principal evaluations, student assessment results, and other relevant information, the superintendent would need a second crystal ball to help determine the rewards and penalties that should be attached to particular performance ranks. In competitive labor markets, efficient innovators thrive and prosper while those who

pursue inefficient personnel policies either abandon them or go out of business, but few competitive forces discipline the personnel policies adopted by nations, states, or even large school districts.

This observation also raises concerns about the ability of large government agencies to determine the reward structures and ranking procedures that govern competition among schools. The benefits of competition among schools will be determined in part by the extent to which policymakers not only choose valid ranking procedures but also attach the correct reward structure to various ranks. Policymakers require enormous amounts of information to perform these tasks well.

Conclusion

The great myth about incentive pay or accountability systems is that they bring business practices or competitive pressures to public education, but such claims are not valid. In contrast to private firms, public school systems are not directly accountable to their customers, that is, the families they serve. In the traditional public school model, teachers and principals, as public employees, are accountable to the appointed agents of elected officials. In accountability or incentive pay systems, teachers and principals are accountable to formulas and procedures created by these same agents. These systems may foster competition to earn the rewards governments offer, but if governments design these competitions poorly, there is no guarantee that they will correct their mistakes in a timely manner.

Decades ago school boards began to adopt policies that guaranteed salary increases for all teachers who obtained master's degrees in education, and our university libraries are now filled with research papers that find no relationship between the acquisition of these degrees and the performance of teachers.²¹ Yet there is no indication that districts intend to break the link between master's degrees and pay levels any time soon. If state education agencies or school districts adopt incentive pay systems that are as ill advised as the decision to grant automatic raises to a teacher who obtains a master's degree, what forces will correct such errors?

Hidden actions of agents can corrupt performance statistics. The multitasking model demonstrates that once government agencies attach important incentives to a particular statistic, government employees will take actions to improve the value of this statistic even if these actions contribute nothing or do harm to those that their organization is intended to serve. However, the political process may corrupt government performance statistics in a more direct manner if interest

DESIGNING INCENTIVE SYSTEMS FOR SCHOOLS

groups exert influence over the adoption of specific performance measures reward schemes for use in incentive pay systems.

The analyses presented here implicitly assume the existence of a benevolent education authority and described the policies this authority might adopt to give its access to information. However, it is easy to imagine ways that ranking procedures and reward structures might be corrupted by the political process. It is inconceivable that an alliance of teachers unions and postsecondary school education could demand that state officials consider the number of master's degrees among faculty members or the total number of hours in professional development classes as a key factor in determining a school's overall performance ranking? It is easy to imagine the adoption of state- or districtwide incentive systems that specify a mapping between certain performance statistics and pay according to rules that do not vary at all with grade, subject taught, or school environment, even though it is almost impossible to justify this approach on efficiency grounds.

These observations suggest that voucher systems and statewide performance measurement systems should be seen not as policy substitutes but rather policies that could work well together. Consider a system that provides comprehensive performance rankings for schools but also allows parents to use information as only one of many factors when deciding where their child should attend school. In this scenario, the choices of parents determine the all budgets of each school, and those who run schools engage in competition for resources by choosing the education and personnel policies that deliver educational services that parents value.

This approach gives parents the opportunity to act on information they that cannot be found in any database and also the opportunity to aggregate information at their disposal based on their values and priorities. By granting parents control over the public funds that are allocated for their children's education, society gains an army of education performance monitors. Lacking control, parents have less incentive to acquire information about the quality of their child's school and no means to credibly communicate the information they possess.²²

Those who are convinced that parents cannot possibly be trusted to choose schools for their children may wish to amend this system by making total school resources dependent not only on student enrollment but also on some government assessment of school quality. But even with such an amendment, a system of real competition among schools may serve as an important catalyst for improving the practices that determine the hiring, retention, and payment of teach-

It is also worth noting that high-powered incentives may not even be the optimal approach to personnel policy in education. The Holmstrom and Milgrom multitasking model points directly to this possibility. In addition, Timothy Besley and Maitreesh Ghatak note that many nonprofit organizations in education, health, or related services choose personnel policies that include relatively little incentive pay. In these types of organizations, they argue, it is often efficient to devote considerable resources to the screening of potential hires and to then hire only candidates with high levels of personal commitment to the mission of the organization. When it is possible to identify such individuals, incentive pay is no longer necessary.²³

Current trends in education reform operate on the assumption that teachers should face high-powered performance incentives, but it is possible that this assumption is wrong. It is possible that, rather than incentive pay systems, schools need much better means for identifying and developing talented persons who enjoy helping children learn. Whether or not this is the case, real competition among schools and organizations that manage schools may be the best mechanism available for societal learning about desirable methods for identifying, training, and motivating teachers.

Notes

1. See chapter 5 in this volume.
2. Edward P. Lazear and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* 89, no. 5 (1981): 841–64, began the economics literature that describes these forms of incentive pay as prizes associated within rank-order performance tournaments.
3. See Donald T. Campbell, "Assessing the Impact of Planned Social Change," *Occasional Working Paper 8* (Hanover, N.H.: Dartmouth College, Public Affairs Center, December 1976).
4. Bengt Holmstrom and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," special issue, *Journal of Law, Economics, and Organization* 7 (January 1991): 24–52.
5. See Eric A. Hanushek and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" Working Paper 10591 (Cambridge, Mass.: National Bureau of Economic Research, June 2004); and Eric A. Hanushek, "Impacts and Implications of State Accountability Systems," in *Within Our Reach*, edited by John E. Chubb (New York: Rowman and Littlefield, 2005).
6. This scenario may be avoided if teachers find that the process of preparing children for specific assessments actually lowers the cost of building other skills. If the process of preparing children for a high-stakes assessment makes it easier to teach critical thinking skills, social skills, and the like, it may be possible to design an accountability system that gener-

ates improved performance on a specific assessment without taking attention and away from the skills that are not directly assessed.

7. Avinash Dixit, "Incentives and Organizations in the Public Sector," *Journal of Human Resources* 37, no. 4 (2002): 696–727, correctly notes that many different advocacy groups act as performance monitors in public education, and these groups do not always have the same priorities. Seen in this light, the typical failure of existing incentive pay systems to take clear and coherent stands on how performance should be defined and measured is completely surprising. However, my goal is not to explain why current government policies are what they are but rather to outline normative criteria that incentive policies should meet.
8. Derek Neal and Diane Whitmore Schanzenbach, "Left Behind by Design: Profit Counts and Test-Based Accountability," University of Chicago, February 2008; Neal and Schanzenbach draw their conclusions based on data from Chicago. Randall Reardon, "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics* 92, nos. 5–6 (2008): 1394–1415, draws similar conclusions based on earlier data from Texas. Matthew G. Springer, "The Influence of NCLB Accountability Plan on the Distribution of Student Test Score Gains," *Economics of Education Review* 27, no. 5 (2008): 556–63, does not find similar patterns using data from Idaho, but he cannot replicate the Neal and Schanzenbach ("Left Behind by Design") methodology because he does not have access to assessments taken before the introduction of NCLB.
9. See Letitia Stein, "Hillsborough's Merit Pay Experiment Benefits Affluent Schools," *St. Petersburg Times*, February 24, 2008 (www.sptimes.com/2008/02/24/Hillsborough/s_merit_.shtml) for details. The new Hillsborough plan is part of the Awards Program that replaced STAR statewide.
10. Chapter 6 in this volume, by Daniel McCaffrey, Bing Han, and J. R. Lockwood, explains in more detail how rankings of teacher performance vary depending on numerous criteria that policymakers must make when building an empirical model to produce the rankings. See State of California, Department of Education, Office of Policy and Evaluation, "Structuring of California's 1999 School Characteristics Index and Similar Schools Ratings," PSAA Technical Report 00-1, April 2000.
11. Hideshi Itoh, "Incentives to Help in Multi-Agent Situations," *Econometrica* 59, no. 3 (1991): 611–36, shows that when the cooperation or helping costs among workers are low relative to benefits, it is optimal for firms to adopt incentives policies that operate only at the team level. In another paper presented to the 2008 conference of the National Center for Performance Incentives, Karthik Muralidharan and Venkatesh Sundararaman, "Teacher Incentives in Developing Countries: Experimental Evidence from India," Working Paper 2008-13 (Nashville, Tenn.: National Center for Performance Incentives, February 2008), using experimental data from India, find no difference in achievement gains associated with teacher incentives as against school incentives. However, the schools involved in their experiment contained only a handful of teachers, and the organization of these schools is greatly different from that of modern schools in developed countries. Gains from cooperation are greatest in larger schools where a number of teachers are teaching similar material to different students. Victor Lavy, "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy* 110, no. 6 (2002): 1286–1300, documents noteworthy responses to a school-level incentive plan in Israel.

13. See Edward P. Lazear, "Pay Equality and Industrial Politics," *Journal of Political Economy* 97, no. 3 (1989): 561–80.
14. The tournament model of Jerry R. Green and Nancy L. Stokey, "A Comparison of Tournaments and Contracts," *Journal of Political Economy* 91, no. 3 (1983): 349–64, clarifies the potential drawbacks of the performance standard approach.
15. Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73, no. 2 (2005): 417–58.
16. See Daniel F. McCaffrey and others, *Evaluating Value-Added Models for Teacher Accountability* (Santa Monica, Calif.: RAND, 2003). for a comprehensive review of value added methods. See chapter 6 in this volume for a detailed case study that explores how variation in methods used to measure teacher effects as well as policies that link reward pay to different performance ranks can, in practice, generate noteworthy variation in distributions of reward pay among teachers. See Daniel F. McCaffrey, Tim R. Sass, and J. R. Lockwood, "The Intertemporal Stability of Teacher Effect Estimates," June 2008, for a detailed treatment of the stability of estimated teacher productivity effects.
17. Jesse Rothstein, "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference," Princeton University, November 20, 2007.
18. Denmark, Netherlands, and Sweden are examples. See Neal (2009) forthcoming for details.
19. In Lazear and Rosen's seminal paper on tournaments, "Rank-Order Tournaments as Optimum Labor Contracts," firms know the exact willingness of workers to supply different levels of effort and the precise relationship between effort and true output, even though neither the worker's contribution to output nor the worker's effort are observed. Similar assumptions are common in many models of bonus pay and promotions. See Canice Prendergast, "The Provision of Incentives in Firms," *Journal of Economic Literature* 37, no. 1 (1999): 7–63.
20. Here, *efficient* does not necessarily mean the first-best outcome in a world with perfect information but rather the best firms can do subject to the information constraints they face.
21. See Kate Walsh and Christopher O. Tracy, "Increasing the Odds: How Good Policies Can Yield Better Teachers," National Council on Teacher Equality, December 2004.
22. Daron Acemoglu, Michael Kremer, and Atif Mian, in "Incentives in Markets, Firms, and Governments," *Journal of Law, Economics, and Organization* (forthcoming), argue that real competition among educators may cause harm. They reach this conclusion because participants in their model are not able to monitor schools directly and thus rely on public statistics like test scores. In this setting, Holmstrom and Milgrom's ("Multitask Principal-Agent Analyses") multitasking model suggests that intense competition among educators may waste resources and harm students.
23. Holmstrom and Milgrom, "Multitask Principal-Agent Analyses"; Timothy Besley and Maitreesh Ghatak, "Competition and Incentives with Motivated Agents," *American Economic Review* 95, no. 3 (June 2005): 616–36.

8

The Performance of Highly Effective Teachers in Different School Environments

William L. Sanders, S. Paul Wright, and Warren E. Langevin

Teacher quality is a major concern of policy leaders and practitioners interested in the condition of American public schooling. A considerable amount of policy debate and media coverage related to issues of teacher quality has focused on schools with large concentrations of economically disadvantaged and minority students. Over the course of their public school education, students in these schools are, on average, not likely to receive instruction of the same quality as students in other schools.¹ This general pattern in the distribution of teacher effectiveness and student outcomes is also being reported with greater frequency in academic journals. If public school students are assigned to classrooms with a disproportionate number of less effective teachers, then the cumulative effect of their lack of exposure to highly effective teachers will likely result in meaningful differences in school attainment and an individual's future earnings.²

In response to concerns over teacher quality, federal and state policymakers have demonstrated a heightened interest in designing practical solutions to motivate highly effective teachers to either move to or remain in high-needs schools.

The working paper on which this chapter is based was supported, in part, by the National Center on Performance Incentives, which is funded by the U.S. Department of Education Institute of Education Sciences (R305A06034). The views in this paper do not necessarily reflect those of sponsoring agencies or affiliated institutions.